

# **METHOD AND STRUCTURE TO IMPROVE THE GATE COUPLING RATIO (GCR) FOR MANUFACTURING A FLASH MEMORY DEVICE**

## **BACKGROUND OF THE INVENTION**

### **1. FIELD OF THE INVENTION**

The present invention relates generally to method and structure for manufacturing a flash memory device. The present invention further relates to the method and structure for manufacturing a flash memory device with enhanced gate coupling ratio (GCR) and thereafter a smaller applied control gate voltage needed in the operation of the flash memory device.

### **2. DESCRIPTION OF THE PRIOR ART**

Nonvolatile flash memory devices such as erasable programmable read-only memory (EPROM) or electrically erasable programmable read-only memory (EEPROM) or flash memory have been widely used and accepted as data storage devices due to their capabilities to store data after the power is turned off.

Generally, a flash memory device is very similar to a metal oxide semiconductor field effect transistor (MOSFET) except that it has a stacked gate structure. Two gates are provided generally in a stacked manner in a flash memory device, one of which is a floating gate, normally formed by polysilicon for the storage of electrical charges, and the other of which is the control gate to control the access of information. The floating gate is generally located underneath the control gate with a dielectric layer (for example,

oxide/nitride/oxide stacked structure) lying between the two gates. The floating gate is named because it is always in the “floating” state without a connection to the exterior circuitry. Instead, the control gate is usually connected to a word line. A gate oxide (sometimes called tunnel oxide) lies between the floating gate and the semiconductor substrate. Additionally, source and drain regions are arranged on the semiconductor substrate on both sides of the gates.

The operation of a flash memory device will be described briefly as follows. During a write operation, a high programming voltage is applied on the control gate, and this forces an inversion region to form in the p-type substrate. The drain voltage is increased to half of the control gate voltage while the source is grounded, and this increases the voltage drop between the drain and the source. In the presence of the inversion region, the current between the drain and the source increases. The resulting high electron flow from the source to the drain increases the kinetic energy of the electrons. This causes electrons to gain enough energy to overcome the silicon/silicon oxide energy barrier and be collected in the floating gate. The floating gate can be erased by grounding the control gate and raising the source voltage to a sufficiently high positive voltage to transfer electrons out of the floating gate to the source gate by tunneling through the thin gate oxide layer.

It is the general trend in the semiconductor market to manufacture devices with miniaturized dimensions, lower operation voltage, lower cost, higher speed, etc. One of the key features of a flash memory device is the gate coupling ratio (GCR) between the floating gate and the control gate, which affects both the operating voltage and device speed. The read/write method of a flash memory is

effected by means of electrons transferring between the floating gate and the source/drain gate.

The gate coupling ratio is defined as the ratio of an induced voltage on the floating gate to the incident voltage applied on the control gate. For the perfect coupling, this ratio is equal to 100%. Generally, increasing the gate coupling ratio can lower the operating voltage and increase the device speed at the same time. A few approaches, which have been taken to increase the gate coupling ratio, include: increasing the overlapped area between the floating gate and the control gate; reducing the dielectric thickness between the floating gate and the control gate; and increasing the dielectric constant ( $k$ ) of the dielectric layer between the floating gate and the control gate.

FIGS.1A through 1H are schematic cross-sectional views showing the progression of steps for forming a flash memory device according to a prior art. As shown in FIG.1A, a semiconductor substrate 110 is provided, the semiconductor substrate has a plurality of shallow trench isolation (STI) structures (not shown) therein. A gate oxide layer 111, a first polysilicon (poly 1) layer 112 and a silicon nitride (SiN) layer 113 are sequentially formed over the substrate 110. The poly 1 layer 112 used as floating gate material is usually grown, for example, by chemical vapor deposition. The silicon nitride layer 113, used as an etching mask for the poly 1 layer etching, is usually also grown by chemical vapor deposition.

As shown in FIG.1B, a photo resist layer 114 is then formed over the top of silicon nitride layer 113 and patterned by a conventional photolithography process. The photo resist layer 114 is used as an etching mask for silicon nitride

layer 113 etching.

As shown in FIG. 1C, an etching process is therefore performed on the silicon nitride layer 113. Furthermore, the photo resist is stripped after the etching  
5 reaction is completed.

As shown in FIG. 1D, a silicon nitride spacer layer 115 is then deposited onto the surface as well as the side wall of the patterned silicon nitride layer 113. Furthermore, in FIG. 1E, an etching process on the poly 1 layer 112 is performed  
10 by using the silicon nitride spacer layer 115 as a self-aligned etching mask. After the self-aligned etching, the silicon nitride layer 113 and the silicon nitride spacer 115 are removed, which is as shown in FIG. 1F.

FIG. 1G shows that an insulating stacked structure 116 is then deposited by  
15 using, for example, chemical vapor deposition. The insulating stacked structure is sometimes called inter-poly dielectric because it is used as the dielectric between the floating gate (poly 1) and the control gate (poly 2).

FIG. 1H shows that a second polysilicon (poly 2) layer 117 is then deposited  
20 and patterned. The poly 2 layer forms the control gate of the flash memory device. After the poly 2 layer etching, drain and source formation (not shown) is therefore conducted by conventional ion implantation process.

In the prior art, the floating gate surface area is the floating gate periphery length times the floating gate width " $L_w$ ". As shown in FIG. 1I, the floating gate is  
25 drawn in a three-dimensional illustration in order to clearly show how the floating gate surface area is calculated. The floating gate height 312 is typically

of thickness 800 Angstroms (A). The floating gate length 314 is typically 2500A. The floating gate width 313 herein is designated as "Lw". Therefore, the floating gate surface area that is effectively coupled to the control gate is the sum of area 311, area 315 and area 316. Therefore,

5                      Floating gate surface area (for the prior art)  
                         =  $(800+2500+800) \times Lw = 4100Lw$

In the prior art as described above, the surface area of the floating gate overlapped with that of the control gate is limited, which results in a gate  
10 coupling ratio of only 60%. It would therefore be desirable to provide a manufacturing method for a flash memory device with a larger surface area of the floating gate overlapped with that of the control gate so that a higher gate coupling ratio can be obtained.

15                      **SUMMARY OF THE INVENTION**

In accordance with the present invention, method and structure are provided for forming a flash memory device with a semiconductor spacer that substantially increases the overlapped area between the floating gate and the  
20 control gate, which in turn results in a substantially improved gate coupling ratio (GCR).

In one preferred embodiment of the present invention, method and structure to improve the gate coupling ratio (GCR) for fabricating a flash memory device are provided. The method and structure include the following steps. A gate oxide  
25 layer, a first semiconductor layer, and an insulating layer are formed sequentially over a provided semiconductor substrate. Then the insulating layer is partially removed until the partial first semiconductor layer is exposed. After

that, a semiconductor spacer is formed on both the exposed insulating layer and first semiconductor layer. The semiconductor spacer is partially removed until the insulating layer is exposed. Then the insulating layer is partially removed in order to expose the first semiconductor layer; thus, the semiconductor spacer protrudes through the surface of the first semiconductor layer. Later an insulating stacked structure is formed over the surface of the first semiconductor layer and the semiconductor spacer. Finally a second semiconductor layer is formed over the insulating stacked structure.

Further objectives and advantages of the present invention will become apparent from a careful reading of the detailed description provided herein below with appropriate reference to the accompanying drawings.

### **BRIEF DESCRIPTION OF THE DRAWINGS**

The foregoing aspects and many of the attendant advantages of this invention will become more readily appreciated and be better understood by reference to the following detailed description, when taken in conjunction with the accompanying drawings, wherein:

FIGS.1A through 1H show the schematic cross-sectional views of the progression of steps for forming a flash memory device in the prior art.

FIG.1I is a three-dimensional illustration showing the floating gate surface area in the prior art.

FIGS.2A through 2I show the schematic cross-sectional views of the

progression of steps for forming a flash memory device in the current invention.

FIG.2J is a three-dimensional illustration showing the floating gate surface area in the current invention.

5

## **DESCRIPTION OF THE PREFERRED EMBODIMENTS**

Reference will be made in detail to the present preferred embodiments of the invention, examples of which are illustrated in the accompanying drawings.

10

FIGS.2A through 2I are schematic cross-sectional views showing the progression of steps for forming a flash memory device according to a preferred embodiment of the current invention. As shown in FIG.2A, a semiconductor substrate 210, typically a silicon substrate, is provided, and the semiconductor substrate has a plurality of shallow trench isolation (STI) structures (not shown) therein. A gate oxide layer 211 (sometimes called tunnel oxide), a first semiconductor layer 212, typically a polysilicon layer, and an insulating layer 213, typically a silicon nitride layer, are sequentially formed over the semiconductor substrate 210. The first semiconductor layer 212, used as floating gate material, is usually grown by chemical vapor deposition. The insulating layer 213, used as an etching mask for the first semiconductor layer etching, is usually grown by chemical vapor deposition.

As shown in FIG.2B, a photo resist layer 220 is then formed over the insulating layer 213 and patterned by a conventional photolithography process.

25

An etching process is then performed on the insulating layer 213. Next, the

photo resist layer 220 is stripped after the etching reaction is completed, as shown in FIG.2C.

As shown in FIG.2D, a partial etching action is then performed on the first semiconductor layer 212 by using the patterned insulating layer 213 as an etching mask. Next, in FIG.2E, a semiconductor spacer 214, for example, a polysilicon spacer, is deposited onto the whole semiconductor substrate surface. Be one choice, the first semiconductor layer 212 is etched entirely until partial gate oxide layer 211 is exposed. Then a semiconductor spacer 214 is deposited over the exposed gate oxide layer 211, the first semiconductor layer 212, and the insulating layer 213. Another choice is to deposit a semiconductor spacer 214 directly over the exposed first semiconductor layer 212 and the insulating layer 213 without etching the first semiconductor layer 212. The semiconductor spacer 214 is used to increase the surface area of the floating gate overlapped with that of the control gate in order to increase the gate coupling ratio (GCR) of a flash memory device. The first semiconductor layer 212 and the semiconductor spacer 214 together form the new floating gate of the current invention and will be designated as layer 230 hereafter.

FIG.2F shows that a further etching process is performed to etch away the top portion of the deposited semiconductor spacer 214 in order to expose the insulating layer 213 lying underneath it. In this etching process, the semiconductor spacer is used as a self-aligned etching mask; in other words, no additional photo masks and photolithography process are needed. Therefore a simpler manufacturing process and lower manufacturing cost are expected by using the method in the current invention. After the semiconductor spacer



etching, the insulating layer is then removed by a conventional etching process, as shown in FIG.2G.

FIG.2H then shows that an insulating stacked structure 215 is deposited by  
5 using, for example, chemical vapor deposition.

FIG.2I shows that a second semiconductor layer 216 is then deposited. The second semiconductor layer forms the control gate of the device. After the second semiconductor layer etching, drain and source formation (not shown) are  
10 therefore accomplished by a conventional ion implantation process.

FIG.2J shows the new surface area of the floating gate in the current invention by a three-dimensional illustration. The new floating gate length is the sum of the length 317, which is typically 1500 Å in the current invention, length  
15 318, and length 319. Lengths 318 and 319 are the semiconductor spacer bottom lengths, typically 500Å each in the current invention. However, the total floating gate length, which is still 2500Å, is not increased in the current invention as compared to that in the prior art. This is very important because the same number of flash memory devices can be fabricated on an 8- or 12-inch silicon  
20 wafer. But the surface area that is effectively coupled to the control gate has been increased to more than twice that in the prior art. The reason will be explained as follows: the new surface area of the floating gate in the current invention is the sum of twice area 320, one of which is on the correspondingly opposite side of the area 320 hidden in this drawing, twice area 321, similarly,  
25 one of which is on the correspondingly opposite side of the area 321 hidden in this drawing, and one area of 322. The height 323 of the first semiconductor

layer is typically 800Å and the semiconductor spacer height protruding above surface 322 is about 1500Å. The new floating gate width, which is still "Lw", is not changed from that in the prior art. Therefore,

$$\begin{aligned} & \text{New floating gate surface area (for the current invention)} \\ 5 \quad & = 2 \times \text{area}_{320} + 2 \times \text{area}_{321} + \text{area}_{322} \end{aligned}$$

According to this calculation, the new floating gate area that is effectively coupled to the control gate is as follows,

$$\begin{aligned} & \text{New floating gate surface area (for the current invention)} \\ 10 \quad & > 9100 Lw \end{aligned}$$

It is now quite clear that the capacitance of the inter-poly layer (the insulating stacked structure) will be increased to more than twice the prior art value since the surface area of the floating gate effectively coupled to the control gate has been increased to more than twice the prior art value in the current invention. The GCR therefore is expected to be increased from 60% in the prior art to more than 75% in the current invention. Once the gate coupling ratio is enhanced, we can therefore expect a lower voltage to be applied on the control gate during device operation. A lower applied control gate voltage further results in less hazard of breakdown of the inter-poly dielectric (i.e., the insulating stacked structure) and hence a better device reliability is expected with the current invention.

Although specific embodiments have been illustrated and described, it will be obvious to those skilled in the art that various modifications may be made without departing from what is intended to be limited solely by the appended

claims.